Web Log File Analysis: Backlinks and Queries[1]

Mike Thelwall

School of Computing and Information Technology, University of Wolverhampton,
Wulfruna Street, Wolverhampton, WV1 1SB, UK.
Email: m.thelwall@wlv.ac.uk
Tel. (01902) 321470

Abstract

As has been described elsewhere, web log files are a useful source of information about visitor site use, navigation behaviour, and, to some extent, demographics. But log files can also reveal the existence of both web pages and search engine queries that are sources of new visitors. This study extracts such information from a single web log file and uses it to illustrate its value, not only to the site owner but also to those interested in investigating the online behaviour of web users.

Introduction

A user visit to a web site will often be recorded as a series of lines in the site's log file. Much valuable information can then be extracted from this about site use and user navigation behaviour[1-3], perhaps using one of the many web tools for summarising web log files, such as NetTracker (www.sane.com). An additional important concern for site owners, addressed in some log analysers, is the generation of new traffic. Web advertisers, for example, put a precise value on this in terms of click-through charges for online banner adverts. Information about how users found a site would be important for those without the brand visibility of, say, The Times newspaper. For this kind of site, then, knowledge of the existence of other pages linking to it and of search engine queries referring visitors are important. The use of links to point to other sites is common and a valid source of study even in the commercial domain because, for example, in the UK external links were carried on 66% of the sites in one survey, where 17% carried links for apparently philanthropic purposes[4].

In the wider social context, the huge web user base makes online activity an important commercial and sociological phenomenon. The study of web surfing behaviour has, however, been very fruitful, particularly in terms of analysing the actions of a user in a single site. Studies of this kind, based upon laboratory experiments or web log analyses, have helped to develop rules for improved site usability. One example of this is the identified need to minimise the number of clicks needed to get to any information from the home page[5,6]. Before a site can be visited, however, it must be found, perhaps through a search portal. Studies of resource discovery through search engines have also been plentiful, covering both evaluations of the tools used and suggestions for improvements[7-11]. It would, however, be useful to know the kind of related analyses that could be performed on site log files, and whether it would be possible to identify any general patterns of user behaviour. One immediate obstacle to this is that log files represent only snapshots of the actions of a set of users. It is contended here that there is, nevertheless, the potential for a close

[1] Thelwall, M. (2001) Web Log File Analysis: Backlinks and Queries, *ASLIB Proceedings*, 53(6) 217-223.

analysis of the data to reveal information that is unexpected in the context of navigation between sites and, therefore, to provide novel hypotheses about aspects of online activity. An example of such a conclusion inferred from log data is the discovery that many users still struggle to understand the web from the appearance of common web site domain names in the top ten of search query terms for 2000[20].

Web Log Studies

Most of the published writings concerning log file analyses of individual web sites have tended to focus on summary statistics or internal navigation analysis[1-3,12-14], or on profiling individual users[15]. This literature has developed general guidelines about the pitfalls in log analysis, particularly concerning the care with which results should be interpreted. Issues of internal site organisation problems discoverable in logs will not be dealt with here because these have been comprehensively covered already. There do not, however, appear to have been any academic articles analysing individual web sites in the context of how they were found from multiple search engines or other sources. Many log file analysers do, nevertheless, have the capability to deliver some or all of the necessary summary statistics, including the SpeedTracer web log analysis tool described by Wu *et al*.[12] which gives a simple count of backlink pages.

There have also been several published exercises that have analysed the logs of the search engines directly, including AltaVista and Excite, with a recent survey article giving an overview of current progress[16]. Such surveys are to some extent inter-web site in nature, dealing with the requests of users seeking to find another site, or information likely to be on another site. The surveys have been statistical analyses and have revealed interesting facts such as average query lengths and the poor take-up of advanced search syntax. They have not, however, attempted to contextualise the information in terms of the site jumped to, focussing instead on information retrieval aspects of the process. There is, therefore, space for some discussion of this point.

The Aims and Scope of the Study

This study takes as its raw data the log files of a single web site over a period of 10 months, covering 17,552 accesses. It reports an attempt to extract as much useful information as possible, in terms of both knowledge useful to the site owner and wider information about web surfing behaviour. This includes an analysis of thousands of search engine queries that resulted in visits and the discovery of pages with links to the site. The first aim is to use a case study to illustrate the kind of information obtainable from the logs of a web site, about how visitors found it. The second aim is to view the data as a snapshot of user web behaviour and to attempt to use it to identify unusual behaviour on the interface between web sites which would merit further research.

Methodology

Identifying Link Source Pages

The web site chosen for the study was the University of Wolverhampton Computer Based Assessment project (cba.scit.wlv.ac.uk). This is quite a small site, with only 5 pages, but one that generates 20,000 page hits per year. It gives information about the project and allows full working programs to be downloaded for free. It's log file could, in theory, contain any information known to the computer system that the web server is operating from, in addition to information sent by the user in the form of web

page requests. For this study the logs contained the text of the request, plus the date and time for each entry. Log information was only kept for web pages and not for the individual graphical elements of the site. Web pages are requested using the HyperText Transfer Protocol (HTTP) and the request must specify, as a minimum, the identifier of the resource sought, normally the portion of the URL after the domain name[17]. Packaged together with the request (although part of a different protocol) must also be details of where to send the page, the Internet Protocol address of the source computer. If this were all the information obtainable then log file analysis would be mainly restricted to site navigation efficacy discussions. Normally, however, additional facts are conveyed with the HTTP header of the request, including the name of the browser or other software package used (HTTP_USER_AGENT) and the domain name of the requesting computer (REMOTE_HOST). One particularly interesting field in the request header is HTTP_REFERER. This, if implemented, can contain the URL of the page from which the current one was accessed. The existence of web pages elsewhere that contain a link to a logged page can be discovered through this information.

Extracting the Search Engine Queries

One type of referrer is especially interesting: a results page from a search engine. This would normally contain an encoded form of the original query submitted by the user. It is particularly useful to give an insight into both the information need that lead the user to the site, and the fact that the search engine's algorithms indicate the site or page as potentially fulfilling that need.

When a web page contains a text box or other HTML form elements and the contents are submitted, they become transformed with a simple encoding and sent to a program residing on a web server, either as part of an URL or separately from that URL[18]. The actual coding of query text in URLs is standard, inherited from their origin in HTML forms, which have a defined procedure for encoding in URLs. Essentially, alphanumeric components of the query copy directly to the URL, but many others change, including spaces to %2E and quotes to %2F. Most search engines use the URL encoding method and so, for example, it is possible to recover queries from the URLs of search engine results pages. There are, in practice, two ways of extracting the query from an URL. The first, available to anyone without the need for specialist software or knowledge, is to copy it to a browser and view the page returned. This may well not contain the same set of links as the original query response, but, unless the search engine no longer supports the form of coding used in the URL, the page should contain somewhere a copy of the original query. The alternative method, used in the survey, is to interpret the URL coding sufficiently to be able to extract the query from it without requesting the page. The factor that is not common to all search engines, however, is the identifier used to distinguish the query text from the other URL-encoded data. These, therefore, had to be manually identified and encapsulated in a program written to extract the query syntax. This variable component, combined with the continual increase in the number of search engines, would quickly render obsolescent any existing web log file analysis programs that attempted to extract search queries. The use of a specially written timely log analysis program was, therefore, essential for a fully comprehensive academic analysis, although the keyword analysis facilities of a standard web log analysis program would be sufficient for normal purposes.

A complication arises when more than one input of information is used to construct a search. Two examples of this are: a search within a search engine

directory category; and an initial search that has been followed by the selection of meta-terms to refine it. Particular care had to be taken to ensure that any extra data of this sort was not lost when programming the extraction of queries.

Results and Discussion

Sources of Page Visits

The log file was processed in order to extract the names of all HTTP referrers. This parameter was present and meaningful in 11,086 out of the 17,552 logged requests. Most of the requests, 3642, were from search engines. Figure 1 shows a breakdown of the 4548 identifiable sources of page visits from external sources, having excluded pages in the CBA site. The 38 visits recorded in the 'Other' category were page requests from translation services, links from bookmarks and 12 visits from a search engine's random jump feature. The graph shows the importance for this site of search engines. It should be pointed out that Yahoo! searches its directory structure first when a query is entered, only searching outside if no matches are found, and so many of the query matches will have been made only because the web site was present in the Yahoo!Directory.

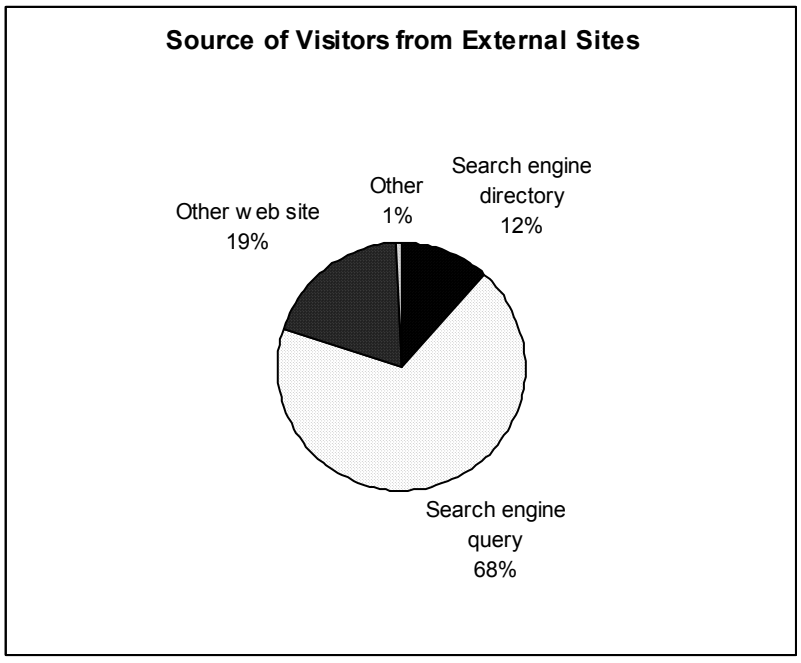

Figure 1. A breakdown of identifiable sources of visitors to the site (from pages outside the host institution) as recorded in the HTTP_REFERER header field of the web request.

Pages Linking to the Site

All search engine request pages and commercial search engine directory pages were filtered out of the list in order to identify other link pages. Of the remainder, 44% came from the university hosting the site. These were also removed, leaving 44 different pages, accounting for 386 site visits in total. This list is useful as a positive indicator of pages that both link to the site and generate traffic to it. The list is not exhaustive, because of the unidentified links. The counts are also not reliable because cache services could also reduce the counts for a page, as visitors after the first through the cache would not register in the site log until the cache flushed the page or requested it again to refresh its archive.

These figures represent, then, the minimum number of site accesses from each source. One important proviso is that the HTTP_REFERER is a parameter that is capable of containing inaccurate information, as the programmer of an unruly browser or crawler could fill it with any text. To counter this, each pages was visited and checked for accuracy. This stage revealed a number of web pages not containing links to the test site. Three were home pages of web sites offering illegal software for free. One more was discovered to be a search engine, which was previously unknown and did not carry the give-away URL-encoded data. A page that stood out as an unusual source of links and appeared 6 times during April and May 2000 with six different identified users, was the home page of the International Atomic Energy Agency, a page that, at the time of checking, did not contain a link to the site. Based upon the six different visitors from the same source, it was assumed that it was likely that a link existed at that time. All of the remaining pages either contained a link or were unobtainable but from their URL seemed likely to contain a link. The final check was of the user agent associated with each visit, which revealed that one of the pages was from a crawler-compiled directory and all of the hits had come from the crawler. This page was discarded.

The cleaned data revealed that a few educational sites were major sources of traffic and a large number of sites or pages generated just a few visits, as shown in figure 2.

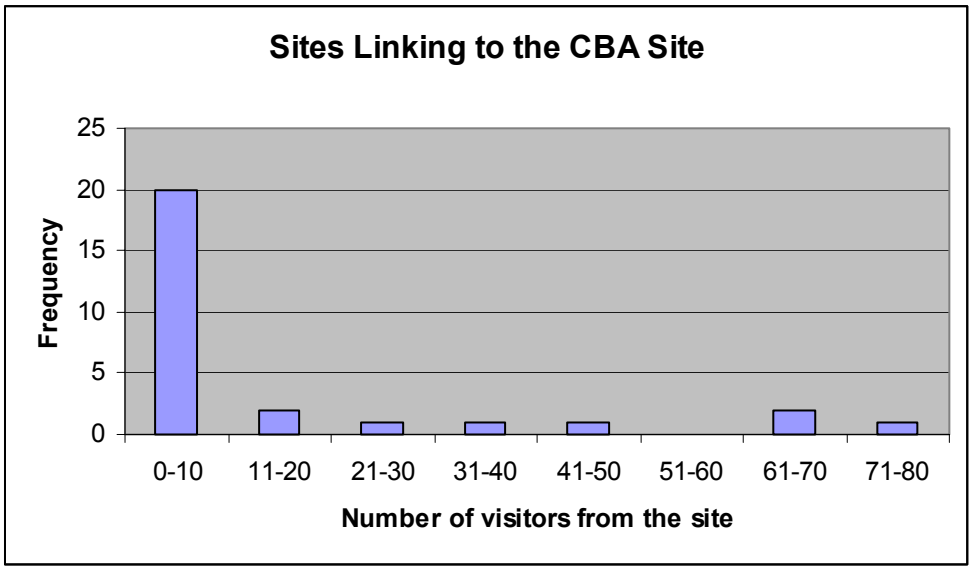

Figure 2. The distribution of external sites in terms of the number of visitors directed to the CBA site.

Comparison of Log File Analysis with Search Engine Backlink Reports

A number of search engines offer the facility to report all of the web pages that link to a given web page or web site. An analysis of the top 16 search engines, as reported by Media Metrix[19] based on a user survey, revealed five offering this facility. A second list of link pages was compiled, this one from combining results from all these search engines: AltaVista; Google; NBCi; HotBot; and Go/Infoseek. For engines that counted links to individual web pages, these were tested for all valid URLs to the five pages in the assessment site, which totalled 14 in all, but almost all links were to the default root domain name of the site only. This list, once search engine directory pages had been removed, was compared with the previous one. The extent of the difference between the two lists was remarkable, as shown in figure 3. Three of the

biggest sources of new visitors were not indexed in any search engine, in both cases because the HTML design effectively hid the link from the search engine. The lack of indexing is interesting because if pages are not indexed in search engines then they must get visitors from other sources. Two of the sites were major educational initiatives, accounting for their use, and one was an academic course page, with the URL presumably communicated to students by their lecturer. No explanation is needed for little-used link pages not being indexed in search engines since they are known to index only a proportion of the web. The existence of pages apparently not generating traffic is also a salutary reminder that being linked to does not guarantee visits. In summary, the combination of two sources provided a more complete picture of the number of pages linking to the site than either offered individually.

All of the pages checked were education related sites linking to the site in order to provide a resource to support education, either targeting those wishing to use the programs or those wishing to see a case study of a computer based assessment project. These two reasons for linking were unsurprising, but it is interesting to note the voyeuristic nature of the latter.

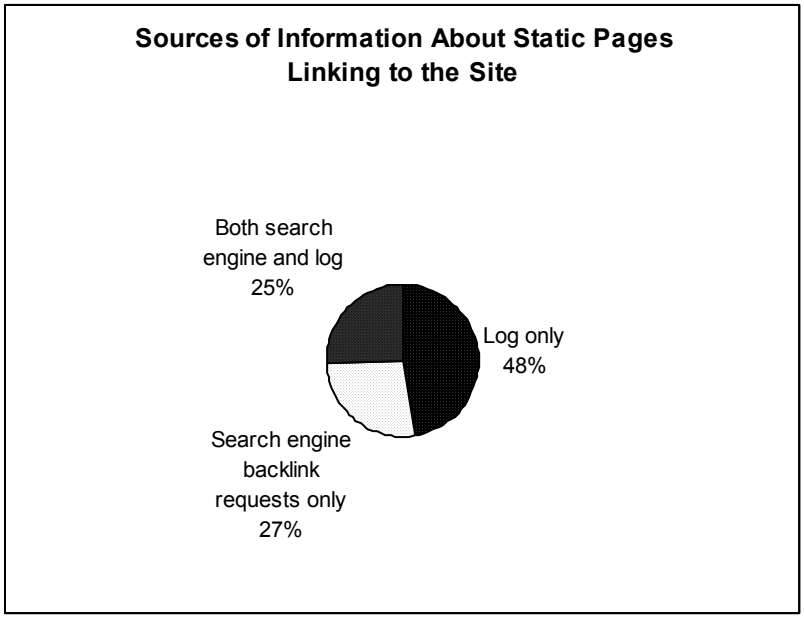

Figure 3. A summary of the proportions of pages found from either and both of the main sources.

Search Engine Queries

As described above, the vast majority of identifiable visitors to the site came from commercial search engines, including their directory. Of these, 56% came from Yahoo!, 11% from Excite, 8% from AltaVista, 6% from Google, and the remaining 22 engines each accounted for 3% or below. Yahoo!'s dominance is a reflection both of its user base, the largest of any search engine during the survey period, and the fact that the site is included in Yahoo!Directory categories, and hence enjoys a relatively privileged position in this search engine.

Of wider interest, however, is an analysis of the query text used in each engine from which a search results page contained a link to the site. These were extracted, grouped and then classified into the nature of the information request, as shown in figure 3. The graph shows that only 16% specifically requested computer based assessment. The largest group requested assessment by using this word or a synonym such as 'test', but did not specify that it had to be computer based or online. Of

course, in some or all of these requests the user may have actually been seeking online resources and may have assumed that this is what they were likely to find since they were searching on the Internet. A small number of searches did not mention assessment, but did include subject-specific details that would allow them to be classified as mathematical, statistical or computing requests. The remainder of the identifiable requests were for purposes outside the objective of the site. A surprising number of visitors searched for Wolverhampton, Wolverhampton University, or some other general request that included the city name. Ten percent also searched for downloadable programs: half for specific programs not available on the site and half for any programs. A small number also used the advanced image search of a search engine to find pictures of a PC or a glider.

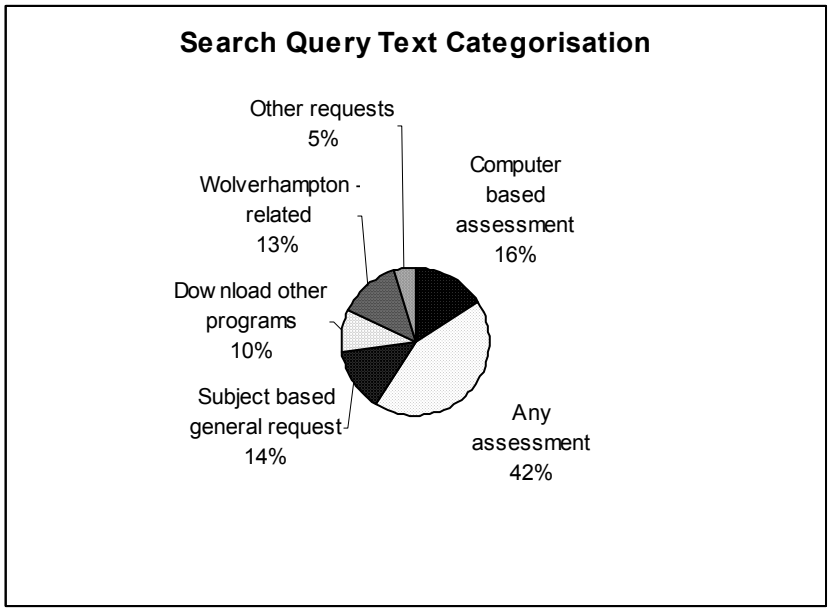

Figure 4. Query text categorisation across all recorded search engine results pages

Conclusion

This study covers only one web site and, therefore, its findings do not necessarily generalise to the whole web. The backlink analysis does, however, suggest several points.

- Search engine backlink counts can be far from exhaustive.
- Many links to a site may generate little or no traffic.
- A combination of search engine backlink discovery facilities and web log analysis is likely to yield a fuller picture of the link structure surrounding a site than either taken on their own.

The existence of unused links is an interesting phenomenon in itself because it is likely to be invisible to the link page owner and not recorded in their logs if the target is an external page. In order to record such information in the log, special programs would have to be connected to the page especially for this purpose. Such is the case, for example, in many search engines, where links to sites are sent via a redirection program, but is believed to be relatively rare on other sites.

The breakdown of search engine queries resulting in new visitors is useful to the owner to identify the search databases that index the site, the keywords perceived as matching its content, and the resulting traffic level. This last indicator would be useful to monitor over time, with falling referrer counts possibly indicating a fall in rank[8]. Of wider interest is the evidence of flexibility of surfing behaviour from those

using a search query and choosing to visit the site. Only 16% included terms unambiguously indicating the need for computer based assessment, although a further 56% expressed a requirement for general assessment or related subject-based information. A fundamental methodological problem here is that it is impossible to infer the precise need which lead to the query formulation. For example, those searching for "maths assessment" may have assumed that the results would be mainly restricted to online tests by virtue of the interrogation medium. Subject to this proviso, however, the queries chosen by visitors were suggestive of at least a flexibility in information needs and perhaps also the ability to be easily distracted from the chosen goal, as evidenced by the 13% who searched for an aspect of Wolverhampton. Clear evidence of parasitic behaviour is also shown in the number of visitors who were looking to download any programs and those simply looking for a general picture on the site.

Web log files, then, can provide useful information to site owners about sources of new visitors. They have also the potential to be a source of evidence to researchers interested in user interaction with search engines as well as more general web surfing behaviour.

References

1. Nicholas, D., Huntington P., Lievesley, N. and Withey, R. 'Cracking the code: Web log analysis.' *Online & CD-ROM Review* 23(5), 1999, 263-269.
2. Nicholas, D., Huntington, P., Williams, P., Lievesley, N. and Dobrowolski, T. and Withey, R. 'Developing and testing methods to determine the use of web sites: case study newspapers.' *ASLIB Proceedings* 51(5), 1999, pp. 144-154.
3. Nicholas, D., Huntington, P., Lievesley, N. and Wasti, A. 'Evaluating consumer Website logs: a case study of The Times/The Sunday Times Website.' *Journal of Information Science* 26(6), 2000, 399-411.
4. Thelwall, M. 'Commercial Web Site Links.' *Internet Research: Electronic Networking Applications and Policy* 11(2), 2001, to appear.
5. Shneiderman, B. 'Designing Information-Abundant Web Sites: Issues and Recommendations.' *International Journal of Human-Computer Studies* 47, 1997, pp. 5-29.
6. Wan, H. A. and Chung, C. 'Web page design and Network Analysis.' *Internet Research: Electronic Networking Applications and Policy* 8, 1998, pp. 115-122.
7. Clarke, S. J. and Willett, P. 'Estimating the recall performance of Web search engines.', *Aslib Proceedings* 49(7), 1997, pp. 184-189.
8. Dowe, D.L., Allison, L. and Pringle, G. 'The Hunter and the Hunted - Modelling the Relationship Between Web Pages and Search Engines.' *Lecture Notes in Artificial Intelligence* 1394, 1998, pp. 380-382.
9. Schwartz, C. 'Web Search Engines.' *Journal of the American Society for Information Science* 49(11), 1998, pp. 973-982.
10. Gordon, M. and Patak, P. 'Finding Information on the World Wide Web: the retrieval effectiveness of Search Engines.' *Information Processing and Management* 35, 1999, pp. 141-180.
11. Oppenheim, C., Morris, A. and McKnight, C. 'The evaluation of WWW search engines.' *Journal of Documentation* 56(2), 2000, pp. 190-211.
12. Wu, K-L., Yu, P. S. and Ballman, A. 'SpeedTracer: a Web usage mining and analysis tool.' *IBM Systems Journal* 37(1), 1998, pp. 89-105.
13. Spiliopoulou, M. 'The laborious way from data mining to web log mining.' *Computer Systems Science and Engineering* 14(2), 1999, 113-26.

14. Spiliopoulou, M. ‘ Web usage mining for web site evaluation.’ *Communications of the ACM* 43(8), 2000, pp. 127-134.
15. Mulvenna, M. D., Anand, S. S. and Buchner, A. G. ‘Personalisation on the net using web mining.’ *Communications of the ACM*, 43(8), 2000, 123-125.
16. Jansen, B. J. and Pooch, U. ‘A review of web searching studies and a framework for future research.’ *Journal of the American Society for Information Science and Technology* 52(3), 2001, pp. 235-246.
17. Fielding, R., Irvine, U. C., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. & Berners-Lee, T. (1999). Hypertext Transfer Protocol -- HTTP/1.1. ftp://ftp.isi.edu/in-notes/rfc2616.txt
18. World Wide Web Consortium, (2001). HTML Home Page. http://www.w3.org/MarkUp/ [Visited 1 March 2001]
19. Media Metrix. ‘Media Metrix Search Engine Ratings.’ http://www.searchenginewatch.com/reports/mediametrix.html [Visited 26 January 2001]
20. ‘Search Engines, Browsers Still Confusing Many Web Users.’ http://cyberatlas.internet.com/big_picture/traffic_patterns/article/0,,5931_588851,00.html [Visited 6 March 2001]